

## Notation

Here we summarize the notations used throughout the survey for the reader’s convenience. Detailed definitions can be found in the reference locations.

Symbol	Description	Reference
$\mathcal{W}$	The world with society systems.	Sec. 1.3.1
$\mathcal{S}$	State space of an environment.	Sec. 1.3.1
$s_t \in \mathcal{S}$	Environment’s state at time $t$ .	Sec. 1.3.1
$\mathcal{O}$	Observation space.	Sec. 1.3.1
$o_t \in \mathcal{O}$	Observation at time $t$ .	Sec. 1.3.1
$\mathcal{A}$	Agent’s action space.	Sec. 1.3.1
$a_t \in \mathcal{A}$	Agent’s action output at time $t$ .	Sec. 1.3.1
$\mathcal{M}$	Mental states space.	Sec. 1.3.1
$M_t \in \mathcal{M}$	Agent’s mental state at time $t$ .	Sec. 1.3.1
$M_t^{\text{mem}}$	<i>Memory</i> component in $M_t$ .	Sec. 1.3.1
$M_t^{\text{wm}}$	<i>World model</i> component in $M_t$ .	Sec. 1.3.1
$M_t^{\text{emo}}$	<i>Emotion</i> component in $M_t$ .	Sec. 1.3.1
$M_t^{\text{goal}}$	<i>Goal</i> component in $M_t$ .	Sec. 1.3.1
$M_t^{\text{rew}}$	<i>Reward/Learning</i> signals in $M_t$ .	Sec. 1.3.1
L	Agent’s learning function.	Sec. 1.3.1
R	Agent’s reasoning function.	Sec. 1.3.1
C	Agent’s cognition function.	Sec. 1.3.1
E	Action execution (effectors).	Sec. 1.3.1
T	Environment transition.	Sec. 1.3.1
$\theta$	Parameters of the world model $M_t^{\text{wm}}$ .	Sec. 12.1.1
$P_\theta$	Predicted data distribution.	Sec. 12.1.1
$P_{\mathcal{W}}$	True data distribution in the real world.	Sec. 12.1.1
$\mathcal{K}$	Space of known data and information.	Sec. 12.1.1
$\mathcal{U}$	Space of unknown data and information.	Sec. 12.1.1
$\mathbf{x}$	Dataset representing scientific knowledge.	Sec. 12.1.1
$\mathbf{x}_{\mathcal{K}}$	Known dataset sampled from $\mathcal{K}$ .	Sec. 12.1.1
$\mathbf{x}_{\mathcal{U}}$	Unknown dataset sampled from $\mathcal{U}$ .	Sec. 12.1.1
$D_0$	KL divergence from $P_{\mathcal{W}}$ to $P_\theta$ at time $t = 0$ .	Sec. 12.1.1
$D_{\mathcal{K}}$	KL divergence from $P_{\mathcal{W}}$ to $P_\theta$ after acquiring knowledge.	Sec. 12.1.1
$IQ_t^{\text{agent}}$	Agent’s intelligence at time $t$ .	Sec. 12.1.1
$\Delta$	Subspace of $\mathcal{U}$ for knowledge expansion.	Sec. 12.1.2
$\mathbf{x}_\Delta$	Dataset from $\Delta$ .	Sec. 12.1.2
$\Theta$	Space of possible world model parameters $\theta$ .	Sec. 12.1.3
$\theta_{\mathcal{K},t}^*$	Optimal world model parameters given the agent’s knowledge at time $t$ .	Sec. 12.1.3
$D_{\mathcal{K},\Theta}^{\min}$	Minimum unknown given the agent’s knowledge and $\Theta$ .	Sec. 12.1.3

Continued on next page

Symbol	Description	Reference
$\mathbf{x}_{1:n}$	Input token sequence.	Sec. 18.1
$\mathbf{y}$	Generated output sequence.	Sec. 18.1
$p$	Probability of generating $\mathbf{y}$ given $\mathbf{x}_{1:n}$ .	Sec. 18.1.1
$\tilde{\mathbf{x}}_{1:n}$	Perturbed input sequence.	Sec. 18.1.1
$\mathcal{R}^*$	Ideal alignment reward (measuring adherence to safety/ethical guidelines).	Sec. 18.1.1
$\mathbf{y}^*$	Jailbreak output induced by perturbations.	Sec. 18.1.1
$\mathcal{A}$	a set of safety/ethical guidelines	Sec. 18.1.1
$\mathcal{T}$	the distribution or set of possible jailbreak instructions.	Sec. 18.1.1
$\mathcal{L}^{adv}$	Jailbreak loss.	Sec. 18.1.1
$\mathbf{p}$	Prompt injected into the original input.	Sec. 18.1.2
$\mathbf{x}'$	Combined (injected) input sequence.	Sec. 18.1.2
$\mathcal{L}^{inject}$	Prompt injection loss.	Sec. 18.1.2
$\mathbf{p}^*$	Optimal injected prompt minimizing $\mathcal{L}^{inject}$ .	Sec. 18.1.2
$\mathcal{P}$	Set of feasible prompt injections.	Sec. 18.1.2
$e_{x_i} \in \mathbb{R}^{d_e}$	Embedding of token $x_i$ in a $d_e$ -dimensional space.	Sec. 18.1.3
$\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$	Projection matrices for query, key, and value.	Sec. 18.1.3
$A_{ij}$	Attention score between tokens $i$ and $j$ .	Sec. 18.1.3
$o_i$	Contextual representation of token $i$ (weighted sum result).	Sec. 18.1.3
$\delta_{x_i}$	Perturbation applied to $e_{x_i}$ , satisfying $\ \delta_{x_i}\  \leq \epsilon$ .	Sec. 18.1.3
$\tilde{e}_{x_i}$	Perturbed token embedding.	Sec. 18.1.3
$A_{ij}^\Delta$	Attention score under perturbation.	Sec. 18.1.3
$\tilde{o}_i$	Updated token representation under perturbation.	Sec. 18.1.3
$\mathcal{H}$	Hallucination metric.	Sec. 18.1.3
$\mathcal{R}$	Actual alignment reward of the model's output.	Sec. 18.1.4
$\Delta_{\text{align}}$	Alignment gap.	Sec. 18.1.4
$\mathcal{L}^{misalign}$	Misalignment loss.	Sec. 18.1.4
$\lambda$	Trade-off parameter for the alignment gap in the misalignment loss.	Sec. 18.1.4
$\mathcal{D}$	Clean training dataset.	Sec. 18.1.5
$\tilde{\mathcal{D}}$	Poisoned training dataset.	Sec. 18.1.5
$\theta$	Model parameters.	Sec. 18.1.5
$\theta^*$	Model parameters learned from the poisoned dataset.	Sec. 18.1.5
$\theta_{\text{clean}}$	Model parameters obtained using the clean dataset.	Sec. 18.1.5
$\Delta_\theta$	Deviation of model parameters due to poisoning.	Sec. 18.1.5
$t$	Backdoor trigger.	Sec. 18.1.5
$\mathcal{B}$	Backdoor success rate.	Sec. 18.1.5
$\mathbb{I}$	Indicator function.	Sec. 18.1.5
$\mathcal{Y}_{\text{malicious}}$	Set of undesirable outputs.	Sec. 18.1.5
$g$	Function estimating the probability that input $\mathbf{x}$ was in the training set, with range $[0, 1]$ .	Sec. 18.2

Continued on next page

Symbol	Description	Reference
$\eta$	Threshold for membership inference.	Sec. 18.2
$\mathbf{x}^*$	Reconstructed training sample in a data extraction attack.	Sec. 18.2
$\mathbf{p}_{sys}$	System prompt defining the agent’s internal guidelines.	Sec. 18.2
$\mathbf{p}_{user}$	User prompt.	Sec. 18.2
$\mathbf{p}^*$	Reconstructed prompt via inversion.	Sec. 18.2